5

This is a Chapter from:
Virtual Body Language
(c) 2011 by Jeffrey Ventrella
published by ETC Press
www.etc.cmu.edu/etcpress
www.virtualbodylanguage.com

A Body Language Alphabet

While helping to get There.com started, I was focused intently on making immersive, fun, and engaging experiences. As more software engineers accumulated in the company, I found it increasingly difficult to communicate with them. Here's why: I was so hell-bent on creating an extraordinary user experience that I was blind to one unavoidable fact: the damn thing had to work *on the internet*. The internet is a messy, confusing place. In fact, it's not a *place* at all. And for real-time animation to work properly and reliably for all participants, no matter where they are physically in the real world, there needs to be a lot of machinery under the hood to make everything appear seamless, natural, and instantaneous. One key reason for this problem is *lag*. In an ideal world, all computers would be able to exchange information as quickly as it takes an impulse to travel from one side of the brain to the other. But Earth is a tad bigger than a brain, and the connectivity is continually morphing. And there are traffic jams that cause *variations* in lag—so it is unpredictable.

Those engineers were doing me, and the company, a favor. They were solving the wickedly difficult problem of building a *distributed simulation*. There is a complex enterprise of real-world message-passing that happens behind the scenes when avatars exchange text, animations, and virtual money. Regarding the components of virtual body language (gestures, facial expressions, poses, moodicons, etc), how do these get packaged up and distributed across the internet? In the last chapter, I described the basic *experiential* components of Avatar-Centric Communication for There.com. But equally important is the design of the underlying system that allows for rapid exchange of mutual body language signals—so the user experience can be *shared*. These signals need to zip through the internet as quickly as possible. Efficient encoding is required.

Components of Body Language

In identifying the basic components of body language that can be implemented in distributed virtual environments, Mania and Chalmers (1998) list six:

1. Facial expressions
2. Gaze
3. Gestures
4. Posture
5. Self-representation
6. Bodily Contact

The more dynamic of these components: facial expressions, gesture, and (shifting) gaze, are my primary focus. They constitute a relatively high-frequency traffic of social signaling, often accompanying the production of text or speech. From a technical perspective, *posture* might be characterized as a slow form of gesture, or a full-body gesture held in place serving as a background over which dynamic gestures are

played. Self-representation is one of the most pervasive forms of body language in virtual worlds like Second Life, where avatar customization is extensive.

The last component in this list—bodily contact—is one of the most powerful and visceral. Bodily contact is problematic in virtual worlds for two reasons: (1) it is technically difficult, due to current state-of-the-art in procedural avatar animation and the simulation of body-to-body physics, (2) it is psychologically and sociologically complicated. When one is immersed in the virtual world, and "one with the avatar", his or her body map is *fused* with the avatar, and the illusion of physical touch is transferred quite strongly.

Consider *griefing*, the anti-social behavior common in multiplayer computer games and virtual worlds whereby the *griefer* annoys, irritates, and harasses others, for no particular reason other than the fact that he can get away with it. Griefing could become especially ugly if there is highly-realistic simulation of physical touch in avatars. Invasion of privacy is already a concern, and the added immersion that physical touch creates could aggravate the problem. This is why the physics collision behaviors of avatars in Second Life (allowing them to bump into each other) can be switched off. This is described rather geekily as the "collisionless avatar setting used for anti-griefing purposes".

But touch isn't necessarily ugly, as we all know. Full-on simulated physical touch can be the basis for some very positive and compelling interactions. And when used by consenting adults—simulated physical touch can be very erotic. But, virtual sex aside (a sprawling topic worth several books), physical touch between avatars can have more common, everyday-like uses, such as handshaking, hand-holding, high-fives, pats on the back, elbow-jabbing, and taps on the shoulder. I will discuss a few of these bodily interactions later. But for now I want to get back on to the subject of a body language alphabet.

Categories of Nonverbal Expression

I identify two general categories of non-verbal expression distributed over the internet: (1) directly-captured expressive motion, and (2) encoded body language. These are analogous to terms that Brian Rotman uses: "non-symbolic media", and "notational media" (2008). Directly-captured expressive motion (non-symbolic media) is read from the communicator's physical body in real-time, either through a video camera, or through motion-capture sensors. The resulting signals are manifested remotely and appear either as streaming video or animated movements on an avatar or some other animated visual form. The movements directly correlate with the movements made by the communicator—possibly with some delay, but *ideally* in real-time. It is the visual equivalent of voice chat—a real-time capture of the communicator's movements.

In contrast, encoded body language (an example of notational media) is visual movement which is encoded and transferred using some kind of *alphabet*. Encoded body language can take the form of a reconstruction of previously-captured gestures and poses, and even abstract visual symbols, blended together to make up composite expressions, much the way a composer might overlap orchestral sounds or the way an artist might overlap image layers in Photoshop. Also, it can be *read* or *played* any time, without the communicator actually being present. It is plastic; it can be manipulated in endless ways (i.e. sped up, slowed down, run backwards, filtered). It is not constrained by the body of the communicator in physical space and time. It is open to cinematic treatment.

Encoded body language is the basis for many highly creative and novel performative art forms, whereby expressive human movement is mediated, amplified, deconstructed, and reconstructed via computer technology. The brilliant interactive Artificial Reality environments of Myron Krueger are but one example of mixing realtime capture with encoded effects. There are many others.

Also significant is the fact that encoded body language can be performed by physically-disabled people. If you want to make your avatar give a very excited, animated presentation while doing a slide show, but the *real you* is sick in bed, you can't be expected to put on a full motion capture suit in order to perform your avatar's high energy body language. And even if you were able and willing to wear a motion capture suit, what happens when you cough, sneeze, or yawn? What if you have an uncontrollable itch in a private part? Wouldn't you prefer to be able to do these mundane, biological acts privately, and not have them projected onto your avatar while you are making a presentation to thousands of people online?

Outering What is Inner

Beethoven was not able to hear any sound (save for a constant "whistle and buzz") when he wrote one of the greatest symphonies of all time. Steven Hawking could not speak or hold a pen when he composed some of the greatest scientific explanations of all time. Likewise, there exist physically disabled character animators who are skilled at virtual body language.

Ludwig van Beethoven

Perhaps the fact that Beethoven had previously been a pianist, and the fact that Hawking had previously been able to speak and write, account for such continued brilliance after their faculties had deteriorated. It is also true of a character animator who is wheel chair-bound but was once able to dance, bound across a rocky creek, or swing on a branch. These artists' brains had acquired the ability to run simulations of reality while their faculties had been intact.

Artistic brilliance can live in the bodymind of a physically-disabled person, and still be manifested in craft, given appropriate alphabetic tools (such as musical notation or a speech synthesizer interface). However, could these forms of virtual expression be performed in real-time? It can take Hawking several minutes to answer a simple question, given his current physical constraints. Could a modern day Beethoven perform music of the complexity and expressivity of the Ninth in a realtime performance? Computer music interfaces are getting more sophisticated all the time. VJs tweak multi-channel visuals in realtime. But what are the physical, cognitive, and technical bounds to such high-bandwidth realtime expression? This question applies to remote body language puppeteering as much as any other performative medium.

I would like to imagine a future master avatar puppeteer who has suffered a debilitating stroke and is paralyzed, but who can still express in a virtual body with the finesse of Buster Keaton, Martha Graham, or Jim Carrey. This would require some kind of encoded realtime body language notation.

Now I'd like to sharpen the focus of these two definitions I have given, explore what encoded body language might look like in the future, and reflect on what that means for avatars.

A Classification

The two categories of non-verbal expression I have just described are both valid and useful. Both will continue to evolve and flourish, because both serve different—and important—purposes for internet communication. Furthermore, they are likely to spawn many hybrid forms, which I'll discuss later. But right now let's delve into these definitions a bit more. Directly-captured expressive motion, while it may require sophisticated technology, has a more straightforward purpose than encoded body language. It must always strive to be as

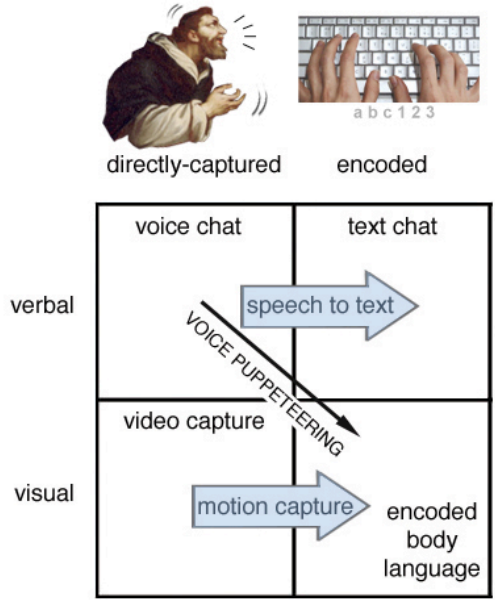
high-res and high-rate as possible so as to broadcast the communicator's realtime energy with verisimilitude. It is not a *language*: it is a *broadcast technology*—with the goal of re-constituting actions happening in a remote location.

Encoded body language, on the other hand, does not have this constraint, for the same reason that the written word is not constrained to the spacetime of its ancestor medium: acoustic sound. But this concept is subtler and harder to define, and still swimming in a hypothetical sea. The beginnings of a unified body language alphabet are emerging. Besides the Facial Action Coding System of Ekman and Friesen, used for simulated facial expression in games and animated films, several researchers are exploring ways to capture body language components and to classify them. This includes systems to automatically annotate body language components using cameras and a computer vision (Chippendale 2006).

Other researchers are investigating language schemes for encoding behavior across a wide range of modalities. The "Gesticon" (gesture lexicon) is a dictionary of behavior descriptions, used by the Behavior Markup Language (BML), described by Kopp et al. (2006). The BML is intended as a standard for multimodal behavior modeling in embodied conversational agents. This idea of the Gesticon has had other proposed names, such as "Gestuary", and "Gestionary" (Rist et al. 2003). Allbeck and Badler (2003) describe a "Parameterized Action Representation" to serve as a linguistic bridge between natural language and virtual agent animation. These are just a few such research projects, many of which are underway today. By the time you read this sentence, there will probably be several more notable research papers on this subject.

The illustration on the next page shows a graph with four modes of communication that I have identified. The top half refers to verbal communication and the bottom half refers to visual communication. The left half refers to directly-captured analog

communication (non-symbolic), and the right half refers to encoded, symbolic, or alphabetical (notational) communication.



A classification of nonverbal communication modalities (Image: Ventrella)

Included in the illustration are two large arrows representing technology for capturing analog human energy and encoding it digitally: speech-to-text, and motion-capture. The diagonal arrow pointing from the voice chat quadrant to the encoded body language quadrant refers to technologies allowing a voice signal to puppeteer an avatar using an automatic gesticulation system. We'll get to this in the chapter, *Voice as Puppeteer*.

The two arrows pointing into the encoded body language quadrant represent two paths for generating encoded body language. But there are many more paths, including the various ways of triggering avatar expressions that I have been describing (through a keyboard, mouse, etc.)

Humanoid Punctuation

Text is now being used widely in conversation—circulating within the realtime interactive loop of language-generation that we find ourselves engaged in while using internet chat systems. New species of punctuation are evolving—almost by necessity, as I described earlier with the Margulian entities. Since the early days of the Medieval reading experience (I'm so glad I wasn't born then), text has acquired more and more forms of punctuation, typographic styles, and visual layout. I like to think of this as part of the ongoing evolution of encoded body language. (These earlier species are the kinds of encoded body language that are meant to be decoded within a reader's mind, so that words can be enlivened with virtual music and motion). But now that text has expanded beyond the static world of ink on paper to something more dynamic and reactive, new layers of expression and nuance are being slathered on.

The early World Wide Web emphasized textual information (and *hypertext*). Today, the internet is becoming increasingly visual and interactive. Flickr and YouTube are two examples of the surge in visual language in the era of social networking. Internet pornography has put a tweak on human sexuality due to its pervasiveness (and its covert use). The internet is becoming more like TV, cinema, games, and…dreams. Is the internet starting to wire itself directly into our limbic systems? Questions of this nature have been explored by several authors (Hayles 1999)(Kastleman 2001)(Carr 2010). The internet may in fact be more amenable to visual (rather than verbal) communication as an ecosystem of virtuality and interactivity. But I'd like to keep the focus on a specific question: given that the internet is becoming more visual and interactive, how shall we characterize the avatar within this scenario?

One view of avatars claims that they originated from computer games. They are descendents of PacMan, and have evolved to a higher dimension. Seen in this light, the avatar is a digital action figure

controlled by a game player. (Note that the characters in The Sims are sometimes called "dolls"). But let's explore the avatar as a *visuo-dynamic text*—a visual counterpart to text communication. Catherine Winters suggested in a blog that "just as virtual environments like Second Life are frequently described as updated MUDs (Multi-User-Dungeons) or chatrooms, user interactions within them can be similarly enhanced by the use of body language and gestures based on that of real-world humans". In this view, the avatar provides embodiment to verbal communication. When considering the avatar as an extension of text chat, it enters the realm of language, a kind of *Humanoid Punctuation*. And since it expresses in realtime, it requires a lot more bandwidth than traditional, static punctuation.

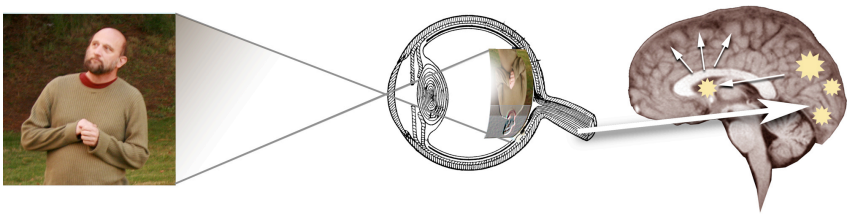
Bandwidth

In order for us to interact intimately with each other over remote distances, the internet will have to take on significantly more data flow. Will the internet be able to physically sustain billions of simultaneous high-resolution video chat conversations? I have asked a few technical people this question, and they say that this will present no problem at all: the internet is well able to handle the load. I remain skeptical, because the human appetite for intimacy is boundless. What happens when we are able to virtualize physical touch?

But even if it is possible for the internet to handle the load, isn't it still a good idea to design more efficient ways to compress communication data?

Consider what is going on in the real world when you are engaged in a conversation with friends around the dinner table. Your eardrums are being vibrated by sound waves within a large range of frequencies (the human ear can detect frequencies ranging from about 20 Hz to about 20,000 Hz—with the human voice ranging between 300 Hz and 3000 Hz). These vibrations are converted into signals in the

frequency domain, and sent down the approximately 30,000 fibers of the cochlear nerve, to the brain. While looking at your friends, billions of photons are landing on your retinas, each of which contain approximately 100 million photoreceptors. Some retinal processing goes on, resulting in signals being sent down the roughly 1.2 million nerve fibers of the optic nerve to the primary visual cortex, after being processed in the thalamus. Signals bounce around, making visits to the various players in the limbic system, and (sometimes) to areas of higher consciousness.



From trillions of photons to body language interpretation (Image: Ventrella)

This is a simplification of what actually happens, but it captures an important notion: we are biologically wired for data compression. There is a progressive reduction of signal density that goes on as these environmental stimuli are filtered, packaged, distilled, and re-represented where pattern recognition and prediction happen, on many levels of cortex. *Attention* helps to filter down this information further, weeding out all but the most relevant signals at any given moment. Over the span of a few seconds, trillions of environmental proto-signals will be received and processed, sometimes resulting in a response as simple as the answer "yes"—a Boolean value.

Or perhaps a nod.

Electromagnetic Expressivity

I have become a regular user of Skype—which allows voice and video calls over the internet. While *Skyping*, I expect to see my wife when we are talking long distance, and I am frustrated when the camera stops working or when the network is slow and choppy, causing my wife to turn into a Francis Bacon painting. Everyone takes telephones for granted, and now a growing wired generation is starting to take video chat for granted. The era of Dick Tracy's wristwatch has arrived—in small mobile devices.

When engaging in directly-captured body language with remote others using wireless devices, electromagnetic radiation is introduced into the sphere surrounding my body. That sphere is where natural body language generation and apprehension normally happens. Whether or not you believe this electromagnetic radiation presents a health hazard (a heavily-debated subject), there is undoubtedly room for optimization to reduce the strength of the signals. Inspiration can be found, I believe, between our ears. We can use the same processes that the brain uses to encode and decode nonverbal signals. Do you know what would happen to your brain if the sum total of raw nonverbal sensory input were NOT reduced to efficient, bite-sized signals? It would fry like bacon. Okay, I exaggerated: but at least the smell of a hardy breakfast might become noticeable.

Natural language favors signals that zip around easily between human minds. Twitter is powerful by virtue of the 140-character limit of tweets, the quantity of tweets, and the rich social contexts within which they live. Poetry, string quartets, advertising brands, popular joke punch lines; all are examples of *memes*, where a lot of meaning is compressed into compact expressions—powerful by virtue of their brevity, mapped against the complexity of the cultural gestalts in

which they thrive. This kind of relationship might be the key to how virtual body language can truly scale on the internet.

Face Caching

Even before I have met someone for the first time, I already have cached in my brain the facilities to recognize a human face—a skill I started to pick up as a newborn. Within seconds of talking to this stranger, I will have cached information about that person's appearances and behavior. As the conversation proceeds, more is cached, and I will need less focus and attention in order to gather the key attributes that make that person unique. I can then start accessing the cached representation of that person in my brain. This is much more efficient than having to continually re-apprehend this person as if I were always seeing him or her for the first time. If I have known this person for many years, I have plenty of cached memory to access, and we can engage using more brief, encoded messaging.

Compare this to an imaging technique called "super resolution reconstruction" which processes a low-resolution image—or most often a series of similar low-resolution images in a video clip—to make a derivative high-resolution image. It can be described briefly like this: imagine a ten second video of a talking head. Any single video frame might have some blur or splotch consisting of only a few pixels, corresponding to a mole or facial wrinkle. The combination of all these blurred images, each slightly shifted because the face has some movement, can be intelligently combined to resolve the detailed image of that mole or wrinkle. The dimension of time is used to determine the resulting appearance in brilliant high-resolution. This is consistent with the way human vision works—within the dimension of time. I would not be surprised if in the future, as I talk to my wife using video chat, the system will have cached attributes of her face, possibly even her body language, much the way I have in my brain. This might come to

pass, not just because it is more efficient and profitable for telecom companies, or because it's a groovy technique, but because it is environmentally friendly…

Local – Global

Think globally, act locally. Here are a few (distant, but perhaps relevant) analogies. Efficient batteries are an important component in stabilizing the energy grid. Locally-stored energy can help offset the volatility of a high-demand stream of immediate-need energy. Mother Nature figured this out a long time ago by making every organism a battery of sorts. Similarly, a move toward agricultural production practices that are more distributed and local can help stabilize the food production process. Being less dependent on large-scale monoculture means less chance of large-scale invasion of pests and contaminations, and the subsequent need for high-tech, environmentally-risky solutions. The current scale of food and energy production and distribution is impacting the global environment. While human communication currently doesn't appear to be having any such effect, that doesn't mean that it wouldn't in the future. Most of us want continual buckets of love, attention, intimacy, friendship, and social intercourse. If human communication continues taking on more technologically-mediated forms—with demands for higher resolutions—then we may be in for more communication traffic jams in the future. And it may even have an environmental impact.

The solution might look something like the solutions to food and energy distribution: offsetting unsustainable centralized systems by making them more distributed and local. In the next several decades, we will see increasing decentralization in all areas, including human interaction. Ray Kurzweil takes this idea to the extreme in *The Singularity is Near*: "The ability to do nearly anything with anyone from anywhere in any virtual-reality environment will make obsolete the

centralized technologies of office buildings and cities" (2005).

Bit Torrent, the peer-to-peer internet file sharing protocol invented by Bram Cohen, is an example of a successful decentralizing system. With Bit Torrent, huge monolithic data files are splintered off into manageable chunks and distributed to several locations. This swarm of chunks organically flows through the internet's irregular terrain, and makes its way to all the locations that requested the data file, reconstituted in-whole. All the participating computers use a common and efficient protocol to gather the data. Something along these lines might be the best way to distribute shared virtual reality.

Shared Reality

Shared reality does not require each participant to have a duplicate copy of the trillion-jillion-gazillion-petabyte database of human experience. Consider that a small handful of phonemes can be combined to generate the words of speech. This is tiny compared to the number of possible words and the number of legitimate ways of combining those words. In linguistics this is called "double articulation". Regarding text, the number of characters, punctuation symbols, words, and rules of usage for their combination is small compared to the number of possible sentences, poems, novels, and books about avatar body language or intestinal health that have been, or could be, written. Borges' Library of Babel is astronomically large compared to any conceivable alphabet.

Sharing an alphabet is what makes it work—for constructing words as well as for generating visual communication. A quick series of subtle overlapping avatar gestures, no matter how complex, can be triggered remotely with just a handful of bytes sent across the wires. As long as both the sender and the receiver have the appropriate code and avatar assets on their local clients, all they need to do is send the right signals between each other to invoke the right gestures, and share

the experience. You can think of the common knowledge and common experiences that human brains share as our model. Like brains, all of our local computers—our viewports into a shared virtual reality—have a set of identical software components cached. Given the right amount of shared context, it would only take a few bits to create a complex experience—identical to all viewers.

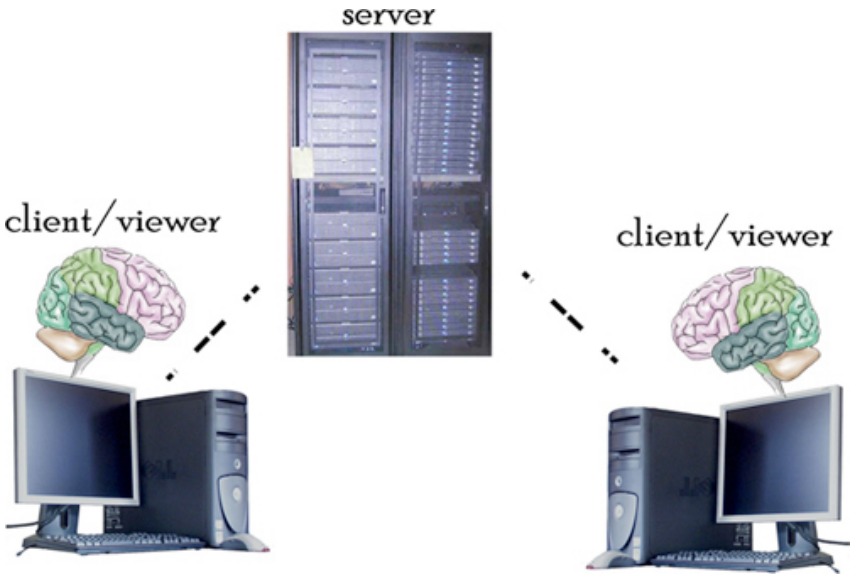
I am reminded of Nicholas Negroponte's oft-quoted reference to the *wink* he expressed to his wife across the table at a dinner party (as I remember him describing it during one of his Media Lab lectures). This wink prompted a knowing grin from his wife because of an experience they had recently shared. All it took was one bit (the wink), and a wash of memories and emotions were evoked. A single wink-bit could also create a connection to the current topic of conversation, placing it in a new context, and conveying an enormous amount of information—100,000 bits-worth. This metaphor has been used to argue that higher network bandwidth is overrated as a necessity (Fiber-optics! Faster speeds! Bigger pipes!)

Wrong: *Smarter* communication.

Two brains with common experience can communicate efficiently (Image: Ventrella)

Local computers that we use to log into current virtual worlds are typically *clients* in a *client-server* architecture. The server acts as a normalizer through which communication happens. And any web developer or virtual world software engineer will tell you: the more

efficient and leaner you can make the data communications over the internet, the better off you are. What better model is there than cognition to use as a guide for making believable, immersive, and expressive virtual worlds?



Clients with shared data can use more efficient communications (Image: Ventrella)

Maturana and Varela's rejection of the "tube metaphor" of communication (1998) is worth mentioning here. According to the tube metaphor, communication is generated at one point (the sender), transmitted through some conduit (like a tube), and then it ends up at a destination point (the receiver). A United States Senator famously described the internet as "a series of tubes" (Stevens 2006). This is not an accurate way to characterize internet ecology—or communication in general. Maturana and Varela prefer a characterization that acknowledges the behavioral dynamic of a whole organic system. Communicative actions may not necessarily follow such clear explicit paths. Also, *information* is not inherent in the stuff that passes through

the tube; it is "created" by the receiver. Similarly, the communicative scenarios I have been describing derive meaning not from what passes through the *tubes of the internet*, but from the way these sweet little efficient packets trigger behavioral changes in a social environment.

Microgestures

Now, back to the body language alphabet. Birdwhistell (who coined the term "kinesics") suggested that the smallest unit of body language should be called the "kine". Referring to the progressive building-up of body language structures, he used terms like "kineme" (analogous to *phoneme*), "kinomorph", and "kinomorpheme" (Birdwhistell 1970). For simplicity, let's just call our elemental units "microgestures". These are atomic, nameable building blocks of body motion that can be combined in series and overlapped in many ways to produce all the familiar non-verbal signals that we use every day.

A leftward cock of the head; eyes glancing upward; eyes glancing to the right; shoulders slumping; a clenched right fist; a slightly-suppressed smile; a stare with eyelids slightly peeled; a step back on the left foot; puffing-out the chest; furrowing the brow; nose-puckering. The list is long, and yet we could probably settle on about a hundred such micro-gestures to treat as *atoms* that can be cobbled together to make exponentially more gestures of more complex non-verbal expression. Even these could be broken down into smaller components. A clenched fist is really a combination of many finger joints bending. It's all a matter of what level of detail we want to use as our base representation.

We may also refer to the facial action components logged by Paul Ekman and Wallace Friesen that can be combined in various ways to make any facial expression (Ekman et al. 1984). In fact, let's incorporate this list into our imagined set of micro-gestures. We could probably fit an enumerated list of well-defined microgestures into a

single byte: 256 possible values—each value serving as an index for a unique microgesture. By combining groups of these microgestures, we could produce almost any imaginable nonverbal expression. The English alphabet has 26 letters. The ASCII standard of character representation adds some punctuation, symbols and control characters, making the set larger. The Unicode standard encoding UTF-8 uses one byte for the ASCII characters, plus up to 4 bytes for other characters. Our body language alphabet could easily fit into a package of a few bytes. Let's throw in a few extra bytes related to timing (i.e., start time plus duration), blending options, and amplitudes. And then let's add a few other parameters determining how the alphabet forms the nonverbal equivalent of words, sentences, and phrases. Sending a few bytes over the internet every second is nothing to blink at these days. And so, it is easy to imagine a compact, efficient protocol for delivering rich body language in real-time.

Of course, this is not the whole picture. If the goal is to build a system that generates meaningful behavior from an encoding, just defining the primitive elements is not enough. Some higher-level filtering, representation, and modeling has to take place. For instance, gestures can be divided into categories, such as those defined by McNeill (1996): *iconics* (gestures that bear resemblance to the thing being talked about), *metaphorics* (describing abstract things), *deictics* (pointing movements referencing spaces, things, concepts, and directions), *cohesives* (gestures that tie together thematically related ideas that become separated in time due to the nature of speech), and *beats* (movements that mark particular words for emphasis). We may consider postural stances, which are often unconscious and can indicate emotion, and *emblems*, which are like visual words (i.e., flipping the middle finger, thumbs-up, salute, etc.) Having higher-level representations like these, as well as established notations, could provide a framework of grammatical rules for piecing together the alphabetical components.

More Rationale for Encoded Body Language

Since humans are able to recognize and process analog communicative gestures, expressions, and postures as semantic units, it seems natural that we should use a similar kind of scheme to deliver body language online. Employing an alphabet and a grammar is more efficient than cramming as many video pixels or 3D model polygons as possible into the pipe as fast as possible to transmit the details of a remote expressive event.

And even though motion-capture techniques can result in data that is more efficient than pixels or polygons, it is still a mere non-linguistic recording of pure movement. Michael Kipp and his research colleagues explain that while these kinds of processes "…result in fine grained, purely descriptive and reliably coded data which can be reproduced easily on a synthetic character, the annotative effort is immense. In addition, the resulting data is hard to interpret. It does not abstract away even minor variations and the amount of data is so massive that it is hard to put it in relation to the accumulated knowledge about gesture structure and form found in the literature". (Kipp et al. 2007) With this in mind, Kipp, and others, have been exploring ways to capture the essential temporal and spatial aspects of conversational gesture and annotate them in a form that can be efficiently re-created on an animated character.

Analog – Digital – Analog

I recently typed the word "Bancouver" into a text document when I meant to type "Vancouver". Since I'm not a touch typist and have to watch the keys most of the time, many typos can go unnoticed. The worst is when I accidentally hit the caps-loCK KEY AND DON'T REALIZE IT'S DOWN UNTIL I LOOK UP AT THE SCREE…damn, there I go again. Regarding "Bancouver", I noticed later that *B* is sitting right next to *V* on my keyboard—a mere centimeter away. This is an

easy mishap when a literary twitch comes out that may have been more productive in the gestural expression space…

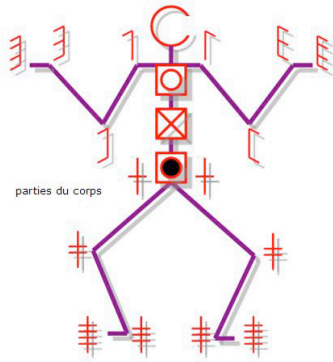
The reason is that gestures are analog. Alphabets are digital. This particular error would not happen in the gestural space of speech and body language. Perhaps a labial twitch could create a *B* sound instead of a *V* sound while discussing Canadian politics over beers (especially several beers). Or perhaps this would be the result of an accent or a language translation. And in terms of bodily expression, a wave of the hand in a northerly direction while referencing Vancouver would not suffer the same kind of "typo" if there were a one-centimeter twitch of a finger.

Animal communication signals can be classified as either *discrete* or *graded*. Discrete signals are either on or off, and are more clearly marked in time. Graded signals can vary over time. They have fuzzier boundaries, and can carry subtle tones of meaning. Discrete signals naturally translate to symbols. Graded signals less so. A rapid uttering of short, discrete sounds over time can result in a subtle, graded-like signal, such as some forms of traditional polyrhythmic African music, the phasing texture of Steve Reich's music, or the electronic music of Aphex Twin (where superfast rhythms blur into gestural phrases).

The 1s and 0s which are the atoms of computing are discrete (as discrete as you can get). When people describe computers in terms of being "just ones and zeros", they are correct only on one level (like, the lowest). That is certainly not what computers are about, any more than you and I are "about" carbon, hydrogen, and oxygen. The signals that computers generate, while digital on the smallest scale, are dedicated to creating analog experiences for us. These signals have reached such blinding speeds and high bandwidths that they create a blur. In virtual world craft, that's intentional. We experience the illusion of analog space and time, movement, and sound—graded signals.

One take-away from this is that all signaling, whether of a discrete or graded nature, can be represented digitally. That makes everything ultimately "alphabetizable", i.e., manipulable as notational media (*including* motion capture).

Notation of human movement is what *Labanotation*—a system originating from the work of dance theorist Rudolf Laban—is centered on: capturing essential aspects of human movement, for choreography, athletics, physical therapy, and many other domains. This general notion of *alphabetizing human movement* brings us back to the ideas of Brian Rotman that we touched upon earlier…



Labanotation (Wikimedia Commons)

The Alphabetic Body

The alphabet and the written word have shaped modern civilization and serve as its dominant *cognitive technology*. But alphabetic texts do not convey the bodily gestures of human speech: the hesitations, silences, and changes of pitch that infuse spoken language with affect. Only to a very small degree do our current handful of punctuations achieve this. Some people believe that speech-to-text processing software is replacing the act of writing as we know it. But why just stop at the actions of our "speech organs" (tongue, teeth, lips…)? Brian Rotman proposes a kind of *gesturo-haptic writing* as a whole-body form of communicating. "…why not notate the movement of any and all the body's organs and parts, aural or otherwise, signifying and a-signifying alike?" He proposes bringing the body back into writing—the "alphabetic body" (2008).

Let's riff on Rotman a bit. Speech-to-text processing software receives the analog sounds of your voice as input, interprets them as a string of words, and types them into a text editor immediately after you utter the words. The sound of the voice—phonetic speech—is an analog signal that is converted into a digital signal (based on an alphabet). Whole-Body Annotation takes this idea to its logical extreme. Again, why just use the voice as input? Video cameras could also be capturing the gestures, facial expressions, and gaze, and laying down a notation parallel to and simultaneous with the alphabetical text being laid down.

Let's use our proposed Encoded Body Language scheme to lay down a body language text, data-compressed for efficient internet communication. This new text layer could be hidden for normal book-like reading of the text. However, the entire text (body and all) could also be "played", using an avatar or some other dynamic visual medium to reconstitute the whole-body expression.

Could this scenario ever be realized? Not sure. The point is: a body language alphabet is reasonable, logical, and possible. We already have a very limited form of body language notation (written text—which codifies phonetic speech). Seeing speech as a limited form of body language allows us to extrapolate to the whole body—and it starts to make more sense. The existence of avatars and the likelihood of their future ubiquity as a *visual text* makes this idea that much more convincing.